J Mol Model (2007) 13:879–890
DOI 10.1007/s00894-007-0197-4

ORIGINAL PAPER

Sequence analysis and rule development of predicting protein stability change upon mutation using decision tree model

Liang-Tsung Huang · M. Michael Gromiha · Shinn-Ying Ho

Received: 25 November 2006 / Accepted: 1 March 2007 / Published online: 30 March 2007

Abstract Understanding the mechanism of the protein stability change is one of the most challenging tasks. Recently, the prediction of protein stability change affected by single point mutations has become an interesting topic in molecular biology. However, it is desirable to further acquire knowledge from large databases to provide new insights into the nature of them. This paper presents an interpretable prediction tree method (named iPTREE-2) that can accurately predict changes of protein stability upon mutations from sequence based information and analyze sequence characteristics from the viewpoint of composition and order. Therefore, iPTREE-2 based on a regression tree algorithm exhibits the ability of finding important factors and developing rules for the purpose of data mining. On a dataset of 1859 different single point mutations from thermodynamic database, ProTherm, iPTREE-2 yields a correlation coefficient of 0.70 between predicted and experimental values. In the task of data mining, detailed analysis of sequences reveals the possibility of the compositional specificity of residues in different ranges of stability change and implies the existence of certain patterns. As building rules, we found that the mutation residues in wild type and in mutant protein play an important role. The present study demonstrates that iPTREE-2 can serve the purpose of predicting protein stability change, especially when one requires more understandable knowledge.

Keywords Bioinformatics · Data mining · Decision trees · Prediction · Protein stability

L.-T. Huang
Institute of Information Engineering
and Computer Science, Feng-Chia University,
Taichung 407, Taiwan

L.-T. Huang
Department of Computer Science
and Information Engineering, Ming-Dao University,
Changhua 523, Taiwan

M. M. Gromiha
Computational Biology Research Center (CBRC),
National Institute of Advanced Industrial
Science and Technology (AIST),
AIST Tokyo Waterfront Bio-IT Research Building,
2-42 Aomi, Koto-ku,
Tokyo 135-0064, Japan

S.-Y. Ho (✉)
Department of Biological Science and Technology,
and Institute of Bioinformatics, National Chiao Tung University,
Hsinchu 300, Taiwan
e-mail: syho@mail.nctu.edu.tw

Introduction

Knowing the relationship between structure, function, and property of proteins is useful to protein design that produces novel protein sequences. Whereas single amino acid mutations can significantly alter the stability of a protein structure, understanding the mechanisms responsible for protein stability change affected by single point mutations has become an interesting topic in molecular biology [1–6]. Until now, various methods have been proposed to predict stability change ($\Delta\Delta G$) upon protein mutation, including energy-based methods and machine learning approaches. Energy-based methods based on force fields can be categorized into three major classes depending on the energy functions [7]: (1) those using physically effective energy functions [8]; (2) those based on statistical potentials for which energies are derived from the frequen-

cies of residue contacts [9, 10]; and (3) those using empirically effective energy functions obtained from experimental data [11]. Lately, machine learning approaches based on artificial neural networks [12] (ANNs) and support vector machines [13, 14] (SVMs) have been proposed. All the above-mentioned methods mainly focused on raising prediction accuracy but not accompanied knowledge acquisition.

The mechanism of systematically and actively capturing knowledge from biological experiment results is valuable to learn an unknown concept. Previously, Xiong et al. [15] have tried to detect repeatedly occurring 3D structures in molecules by finding patterns in 3D graphs. Besides, association rules which demonstrate diverse mutations and chemical treatments have been reported using the A priori algorithm from 300 gene expression profiles of yeast [16]. Moreover Oyama et al. proposed a data mining method to discover association rules related to protein-protein interactions [17].

Biological data mining is one of the emerging research topics in bioinformatics [18], which can be applied in several aspects including description, estimation, prediction, classification, clustering and association [19]. For further understanding the connection between protein sequences and stability change, data mining techniques such as sequence analysis and rule development are necessary. Although tertiary structure information has been used to predict stability change [12–14, 20] and non-local interactions are the principal determinant of protein stability [9], the relevant information may be unavailable. Besides, mutagenesis experiments may begin from the proteome in post-genomic era, developing prediction methods based on sequence information is considerably necessary. Moreover, previous investigations also revealed that local interactions and primary sequence information can play important roles in stability prediction [9, 21] and have introduced them into their works effectively [9, 21–26].

The integration of knowledge acquisition and building interpretable prediction would provide deep insights to understand the mechanism for protein stability as well as to predict the protein stability change upon mutation. In our earlier work (iPTREE), we have proposed a classification tree algorithm for discriminating the stabilizing and destabilizing mutants with high accuracy [20]. The performance of the method is better than or comparable with other machine learning [artificial neural networks (ANNs) and support vector machines (SVMs)], and energy-based methods [12, 13]. In this work, we developed an interpretable prediction method (named iPTREE-2) based on a regression tree algorithm, which aims to achieve the following two goals: (1) accurately predicting change values of protein stability upon single point mutations from sequence information; and (2) simultaneously analyzing protein sequences and developing interpretable rules for knowledge acquisition of stability change by iPTREE-2.

Materials and methods

Protein and mutant datasets

For comparisons, the dataset used by Capriotti et al. [13] was trained and tested in this study. The dataset was obtained from ProTherm database (http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html) [27, 28]. The constraints of extracting the dataset of proteins from ProTherm were: (1) only single point mutations (no multiple mutations) were considered for each protein; (2) the correspondent energy changes of protein stability were detected by experiments; and (3) experimental conditions of temperature and pH were listed in the database as well. And therefore we have accordingly acquired the whole dataset from http://gpcr.biocomp.unibo.it/~emidio/I-Mutant2.0/dbMut.html. The dataset consists of 2048 different single point mutations and is comprised of 64 protein sequences.

The dataset suffers from a great deal of redundancy. To avoid any redundancy bias, we removed 63 redundant samples that have the same experimental setting and $\Delta\Delta G$ values as some other sample. In addition, there are samples that have the same experimental setting and very highly correlated $\Delta\Delta G$ values, which might be under different salt conditions or buffers, ions etc. By taking the average of those samples, another 126 samples were removed in this case. We refer to this redundancy-reduced dataset as S1859.

The proposed iPTREE-2

iPTREE-2 based on classification and regression tree algorithm (CART) can provide a sequence based input with sequence analysis, and then build a prediction model from which interpretable rules can be developed effectively. Figure 1 illustrates the framework of iPTREE-2, where both paths beginning from the input dataset block to the sequence analysis block and to the rule development block are essential. The first path shows the statistical examination of input sequences from the viewpoint of composition and pattern. The second path depicts the process of building the prediction model and developing the rules which is helpful to explore the hidden information in datasets. The major parts of iPTREE-2 are described as follows.

iPTREE-2 kernel- classification and regression tree algorithm (CART)

CART [29] is a machine learning technique which is capable of solving classification and regression problems

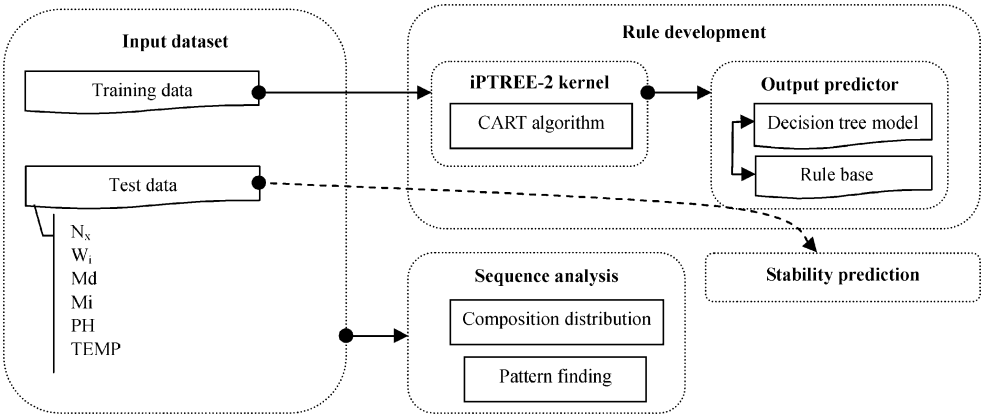

Fig. 1 The framework of iPTREE-2 for rule development and sequence analysis of protein stability change prediction

for both categorical and continuous dependent variables. The main advantages of CART include: (1) being nonparametric to suit data with unknown as well as skewed distribution, no assumptions are required or made regarding the underlying distribution of independent variables; (2) being robust to handle extreme values by isolating the outliers in a separate node, the negative effect of noisy data that may affect model building in addition to accuracy can be reduced; and (3) being efficient to search all possible variables as splitters, problems with hundreds of possible independent variables can be overcame. Due to those abilities mentioned above, CART has successfully been applied for the prediction in molecular biology such as absorption classes [30], the passage of molecules etc [31].

CART mainly consists of two steps, during which one tree is built and then pruned. In the first step, a recursive split procedure builds a tree, named maximum tree, which closely describes the training dataset. In the second step, the maximum tree is cut off for finding optimal subtree. The details of these steps are described below.

Building maximum tree CART constructs a binary decision tree based on training dataset, starting at the tree root [29]. The dataset will be progressively split into smaller subsets which satisfy a given condition. The splitting procedure is made in accordance with squared residuals minimization criterion which implies that expected sum variances for two resulting nodes should be minimized:

$$\arg\min_{x_i \le x_i^{\mathrm{R}}, i=1,\ldots,\mathrm{M}} [P_l \mathrm{Var}(Y_l) + P_r \mathrm{Var}(Y_r)] \quad (1)$$

where P_l, P_r are fractions of samples in the left and right nodes; Var(Y_l), Var(Y_r) are variances of response vectors for corresponding left and right child nodes; $x_i \le x_i^{\mathrm{R}}$ is the optimal splitting condition which satisfies the criterion (Equation 1) with x_j^R the best splitting value of variable x_j from M variables in learning samples. Each available variable is evaluated using squared residuals minimization criterion, and the best variable is selected and used to split samples. The training samples are divided to appropriate descendant nodes which were arranged in alphabetical order of variables. The process is then repeated using the training samples associated with each descendant node.

The splitting procedure goes on until: (1) only one observation or more than two observations with the identical values exist in each of the child nodes, or (2) the number of levels exceeds the limit set by system. Then the procedure of building the maximal tree is terminated.

Tree pruning The maximum tree may turn out to consist of hundreds of levels with high complexity and accompany overfitting which is a considerable problem for many learning algorithms. In order to ease the difficulties, one approach to avoiding overfitting in practice is tree pruning [29] which implies choosing the right size of tree, namely, cutting off insignificant nodes. In practice, there are different pruning algorithms which can be used to choose the right size of tree. However, we have tried several ones and found no one with a significantly superior performance. Thus, instead, we list and observe the results obtained by different pruning levels.

For finely tuning the level of tree pruning, iPTREE-2 made use of a complexity parameter, denoted α, which is a penalty for additional nodes. Usually, as α increased, more and more nodes are pruned away, resulting in simpler and simpler subtrees. For the maximum tree T, the subtree T_α was built by pruning the number of nodes down to $|T_\alpha|$. It is defined as

$$|T_\alpha| = (1-\alpha)|T| \quad (2)$$

where α is from 0 to 1, $|T_\alpha|$ and |T| are the total number of nodes in the subtree and the maximum tree, respectively. The removed nodes are designated in opposite order of tree construction. The appropriate value of α is problem-dependent. "Prediction from sequence composition and tree pruning" section shows that the parameter α directly affects prediction performance of iPTREE-2.

Input dataset and sequence analysis

The ability of processing different variable types in CART makes it flexible to varied applications. Based on the sequence information of S1859, iPTREE-2 can utilize the following independent variables as input:

(1) N_x, the number of the encoded residue type X, is found inside a symmetrical window centered at the mutated residue, namely, the relative abundance of the corresponding residue of neighbors in an extended sequence towards the left (N-terminus) and the right (C-terminus).
(2) W_i represents the sequential neighbors of the mutated residue in the window, where i labels the position of neighbors in the window from N-terminus to C-terminus. For example, when the window size 7 is considered, three neighboring residues on both sides of the mutant residue are labeled with W_1, W_2, W_3 (from N-terminus) and W_4, W_5, W_6 (toward C-terminus), respectively.
(3) Md means the mutation type of deleted-residue or mutation residue in wild type.
(4) Mi means the mutation type of introduced-residue or mutation residue in mutant protein.
(5) PH stands for the pH value of the experimental condition.
(6) TEMP stands for the temperature (°C) at which the stability of the mutated protein was measured explicitly.

In order to further demonstrate iPTREE-2 in various aspects, different combinations of variables were designed. For comparison with previous SVM-based predictor, the same variable set C-19 includes Md, Mi, PH, TEMP and N_X that was obtained from window size of 19 residues. Next, for observing the possible effects caused by different window sizes, N_X in C-19 was re-calculated for varied window sizes 7, 11, 37 and 55, producing another three variable sets C-7, C-11, C-37 and C-55, respectively (see "Prediction from sequence composition and tree pruning" section). Mainly, the sequence information is composed of composition and order of sequences. For further clarifying the role of the two items in stability prediction, additional variable sets named O-7 and CO-7 were used with C-7:

(1) C-7 includes Md, Mi, PH, TEMP and N_X that was obtained from window size of 7 residues;
(2) O-7 replaces N_X of C-7 with W_i; and
(3) CO-7 adds W_i into C-7.

The three variable sets were designed on one-factor-at-once strategy to compare the relative importance of those variables which are different parts between any two variable sets. In the similar approach, variable sets with window size 55 were designed to be compared (see "Prediction from sequence order" section).

Sequence analysis, one of fundamental data mining techniques, was applied to S1859 in terms of sequence specificity, including composition and order of residues. On the one hand, for compositional specificity, the relative abundance of each residue type of neighbors was calculated for different $\Delta\Delta G$ ranges. Meanwhile, in each $\Delta\Delta G$ range, the percentage of each residue type was calculated (see "Sequence analysis of residue composition distribution" section). On the other hand, triplets and quintets of sequences with high frequency of occurrence are found for order specificity. Subsequently, the number of times they happen in different $\Delta\Delta G$ ranges was also observed (see "Sequence analysis of pattern finding" section).

Output predictor

Base on S1859, iPTREE-2 can build a learned decision tree model to be used to predict change values of protein stability. The model can bring forth important factors and be developed into a rule base, for the purpose of data mining. iPTREE-2 relies on a greedy search which iteratively selects the candidate that minimizes a heuristic splitting criterion from input variables. The order of selection will expose the contribution of variables to predict stability change. Usually, the earlier the variable is selected, the more important it would be.

Moreover, the learned decision tree can transform to a rule base that consists of decision rules. Each leaf node has a decision rule, the conjunction of the decisions leading from the root node through the tree to that leaf. For a decision rule, the redundant antecedents can be deleted to simplify the rule. The interpretable knowledge gives biochemistry experts the possibility of validating the predictor and making a discovery (see "Exploring important factors and decision rules" section). By contrast, investigators who want to obtain more information by many other predictors would suffer a setback. Take ANNs for example, describing as well as analyzing the interaction relationship between neurons are extremely difficult.

Prediction scores and test procedures

Several scoring functions were used to decide whether the values between prediction and experiment value can fit well or not. Besides, two test procedures were applied to assess the validity of the results

Single correlation and standard deviation

Pearson product-moment correlation coefficient (R) was adopted to evaluate relationship between stability change value in experiment and in prediction:

$$R = S_{XY}/S_X S_Y \quad (3)$$

where S_{XY} is covariance for variable X and Y, described as follows:

$$S_{XY} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \quad (4)$$

where X_i and Y_i is the stability change in experiment and in prediction, respectively; $\overline{X}$ and $\overline{Y}$ is the mean of X_i and Y_i, respectively. N is the total number of mutants. The standard deviation (STD) of a variable Z_i is the square root of the sum of the squared differences around the mean divide by the sample size:

$$\mathrm{STD} = \sqrt{\frac{\sum_{i=1}^{\mathrm{N}} \left(Z_i - \overline{Z}\right)^2}{\mathrm{N}}} \quad (5)$$

where $\overline{Z}$ is the mean of Z_i.

Mean absolute error

The mean absolute error (MAE) is defined as the absolute difference between predicted and experimental stability values:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \left|V_i^P - V_i^E\right| \quad (6)$$

where V_i^P and V_i^E are the predicted and experimental stability values respectively, N being the total number of mutants.

Self-consistency and n-fold cross-validation tests

The present method was validated by both self-consistency and *n*-fold cross-validation tests. Self-consistency applies all the mutant of dataset to train iPTREE-2 and the prediction is made to themselves. To avoid the requirement for a new or independent validation dataset, *n*-fold cross-validation partitions samples into *n* sub-samples chosen randomly with approximately equal size. For each sub-sample, iPTREE-2 builds a tree model from the remaining data and uses it to predict the stability of the sub-sample. Then the procedure repeats for n times to obtain the related scores.

Results and discussions

Prediction from sequence composition and tree pruning

In Table 1, we present the comparison of correlation coefficients resulted from different window lengths and pruning levels using iPTREE-2. The data were obtained with the dataset of S1859 and 20-fold cross-validation test. To avoid the overfitting problem, we also manipulated pruning level by the parameter α that varied from 0 to 0.5 since too large α value results in huge cutting to reduce the accuracy. The computing platform is Intel Celeron processor 2.4 GHz with 768 MB RAM running Microsoft Windows XP.

Table 1 Comparison of correlation coefficients* resulted from variable sets of different window sizes of sequence composition by using varied pruning levels

α	Variable set				
	C-7	C-11	C-19	C-37	C-55
0.00	**0.66**	0.64	**0.67**	**0.66**	0.69
0.05	0.63	0.64	0.66	0.64	0.67
0.10	0.65	0.64	0.67	0.65	0.67
0.15	0.64	**0.65**	0.67	0.65	0.69
0.20	0.65	0.63	0.66	0.65	0.68
0.25	0.65	0.63	0.67	0.66	0.69
0.30	0.63	0.65	0.65	0.65	0.69
0.35	0.65	0.65	0.66	0.64	**0.70**
0.40	0.64	0.64	0.65	0.65	0.69
0.45	0.65	0.64	0.65	0.63	0.68
0.50	0.64	0.64	0.64	0.66	0.67
Mean	0.65	0.64	0.66	0.65	0.69
STD	0.009	0.007	0.009	0.008	0.009

* obtained with 20-fold cross-validation test.
STD: standard deviation

For understanding the performance difference between iPTREE-2 and the previous method that has predicted the value of stability change on the dataset, the same variable set C-19 with composition information was used. When the tree model was built with no pruning (α=0), we obtained a correlation of 0.67 between predicted and experimental values. The result is significantly better than that obtained from the previous SVM-based method [13] using sequence information (R=0.62) on a redundant dataset, on which iPTREE-2 showed a correlation coefficient 0.73. When performing tree pruning, it shows tree pruning can affect the performance of iPTREE-2. However, the insignificant improvement between the highest and lowest correlation coefficients reveals that the effect is slight.

For different window sizes of sequence composition, the mean correlation coefficient with size 55 (R=0.69) is higher than that with size 7 (R=0.65). It seems that the larger window size can enhance the prediction performance. Interestingly, the α values for the highest correlation coefficient (label with bold type) of each variable set spread between 0 and 0.35, which implies that tiny tree pruning helps prediction performance. As a result, the highest correlation coefficient (R=0.70) ,which conforms to the observations on pruning level and window size, appears

on set C-55 with α=0.35. The result is also comparable with that obtained from the previous SVM-based method [14] using a smaller dataset. Figure 2 shows the scatter diagram with correlation coefficient 0.70 between experimental and predicted ΔΔG values to observe their relationship. Detailed analysis of this figure revealed that the stability of 67% of the mutants is predicted within the error of 1 kcal mol^{-1}. Further, we noticed that 30 mutants (1.63%) are outliers and the removal of these mutants improved the correlation up to 0.75.

Prediction from sequence order

This section introduces the information of sequence order to lead to a discussion about relative importance between two factors, composition and order, to the prediction of stability change. The designed variable sets, C-7 and O-7, can be respectively regarded as the composition based input and the order based input, then CO-7 represents the combination of both. In Table 2, the mean correlation coefficient of O-7 is slightly higher than that of C-7 (R=0.66 and R=0.65 respectively), but not lower than that of CO-7 (R=0.66), which indicates that the order based input is superior to the composition based input on the small scale of window size of 7. And the input of combination of composition and order information can not further improve the prediction performance.

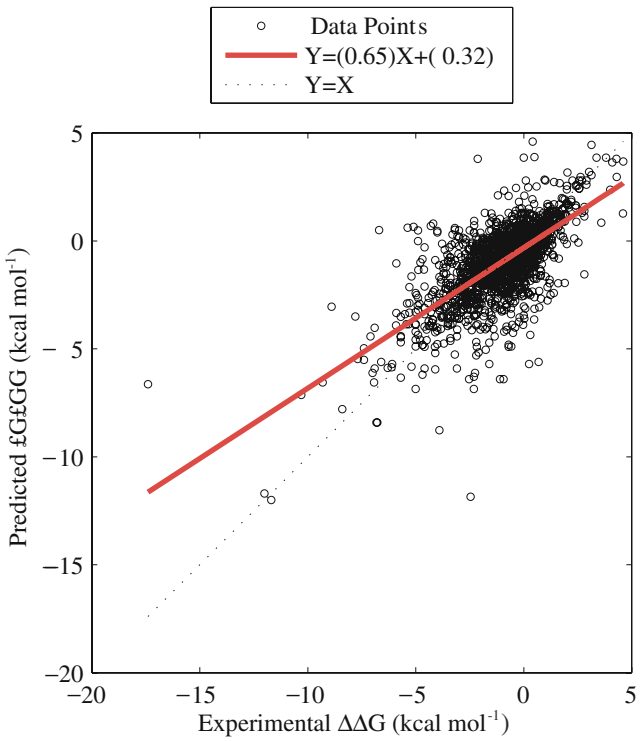

Fig. 2 The scatter diagram for the dataset S1859 of experimental and predicted ΔΔG values using iPTREE-2 with 20-fold cross-validation test

Table 2 Comparison of correlation coefficients* resulted from variable sets of sequence composition and order by using varied pruning levels

α	Variable set					
	C-7	O-7	CO-7	C-55	O-55	CO-55
0.00	**0.66**	0.66	0.67	0.69	0.67	0.66
0.05	0.63	**0.67**	0.67	0.67	**0.70**	**0.68**
0.10	0.65	0.64	0.67	0.67	0.67	0.66
0.15	0.64	0.65	0.66	0.69	0.67	0.66
0.20	0.65	0.67	0.65	0.68	0.67	0.66
0.25	0.65	0.66	**0.68**	0.69	0.66	0.68
0.30	0.63	0.66	0.67	0.69	0.66	0.65
0.35	0.65	0.65	0.67	**0.70**	0.65	0.67
0.40	0.64	0.67	0.66	0.69	0.67	0.67
0.45	0.65	0.67	0.67	0.68	0.66	0.68
0.50	0.64	0.66	0.65	0.67	0.65	0.67
Mean	0.65	0.66	0.66	0.69	0.67	0.67
STD	0.009	0.008	0.009	0.009	0.01	0.009

* obtained with 20-fold cross-validation test.
STD: standard deviation

Comparing with those on a larger scale of window size of 55, the composition based input results in a higher correlation (R=0.69) than the order based input (R=0.67), while the combination of both can not make additional progress (R=0.67). One explanation for this is that the composition of larger scale of widow size may contain relative important prediction information that covers enough of the order information, so that the prediction performance can not be risen after adding the order information. We also list the results on CO-7 by self-consistency test in Table 2. Overall, when based on sequence information including both composition and order, selecting a small window size as input is sufficient and efficient. When based on sequence information including composition without order, choosing a larger window size is helpful to lift up the prediction performance.

Sequence analysis of residue composition distribution

The residue composition of sequences has been shown the importance of improving prediction performance from above. In order to explore the compositional specificity, the frequency distribution of the value of stability change for 20 types of residues was calculated. Tables 3 and 4 was calculated from sequences of window size 7 and 55, respectively. The numbers in parentheses refer to the relative percentage of residue occurrences for each stability change range. We then labeled the relative percentages which are higher than or equal to 10 percent in bold type to emphasize the importance of corresponding residue types.

Table 3 Frequency and percentage distribution of the value of stability change for 20 types of residues from window size 7

Residue	ΔΔG range (kcal mol^{-1})													
	>=3.0	2.5∼3.0	2.0∼2.5	1.5∼2.0	1.0∼1.5	0.5∼1.0	0.0∼0.5	(−0.5) ∼ 0.0	(−1.0) ∼ (−0.5)	(−1.5) ∼ (−1.0)	(−2.0) ∼ (−1.5)	(−2.5) ∼ (−2.0)	(−3.0) ∼ (−2.5)	<(−3.0)
A	4	9	8	24	34	97	223	200	139	87	55	45	35	105
	(4.8)	**(12.2)**	(6.3)	**(11.4)**	(6.9)	**(10.9)**	**(14.0)**	**(11.4)**	**(10.7)**	(7.2)	(6.2)	(6.2)	(7.3)	(8.9)
R	0	1	4	12	21	48	90	101	61	75	50	42	21	68
	(0.0)	(1.4)	(3.2)	(5.7)	(4.3)	(5.4)	(5.6)	(5.7)	(4.7)	(6.2)	(5.6)	(5.8)	(4.4)	(5.8)
N	0	2	7	5	21	33	79	98	53	42	51	50	28	61
	(0.0)	(2.7)	(5.6)	(2.4)	(4.3)	(3.7)	(4.9)	(5.6)	(4.1)	(3.5)	(5.7)	(6.9)	(5.8)	(5.2)
D	4	4	5	12	21	39	89	94	77	96	59	45	32	75
	(4.8)	(5.4)	(4.0)	(5.7)	(4.3)	(4.4)	(5.6)	(5.3)	(5.9)	(8.0)	(6.6)	(6.2)	(6.7)	(6.4)
C	5	2	5	7	21	21	31	56	30	27	17	18	9	23
	(6.0)	(2.7)	(4.0)	(3.3)	(4.3)	(2.4)	(1.9)	(3.2)	(2.3)	(2.2)	(1.9)	(2.5)	(1.9)	(2.0)
Q	2	4	9	14	21	36	53	72	60	48	30	26	18	47
	(2.4)	(5.4)	(7.1)	(6.7)	(4.3)	(4.1)	(3.3)	(4.1)	(4.6)	(4.0)	(3.4)	(3.6)	(3.7)	(4.0)
E	7	4	8	15	46	62	95	94	87	81	74	44	34	48
	(8.3)	(5.4)	(6.3)	(7.1)	(9.4)	(7.0)	(6.0)	(5.3)	(6.7)	(6.7)	(8.3)	(6.0)	(7.1)	(4.1)
G	1	3	7	10	30	74	127	134	110	101	57	46	26	72
	(1.2)	(4.1)	(5.6)	(4.8)	(6.1)	(8.3)	(8.0)	(7.6)	(8.4)	(8.4)	(6.4)	(6.3)	(5.4)	(6.1)
H	0	0	0	1	6	7	27	15	10	18	16	14	8	17
	(0.0)	(0.0)	(0.0)	(0.5)	(1.2)	(0.8)	(1.7)	(0.9)	(0.8)	(1.5)	(1.8)	(1.9)	(1.7)	(1.4)
I	5	4	7	12	14	38	63	83	61	73	45	53	26	69
	(6.0)	(5.4)	(5.6)	(5.7)	(2.9)	(4.3)	(3.9)	(4.7)	(4.7)	(6.1)	(5.1)	(7.3)	(5.4)	(5.9)
L	0	3	16	30	52	92	150	166	108	79	62	48	42	111
	(0.0)	(4.1)	**(12.7)**	**(14.3)**	**(10.6)**	**(10.4)**	(9.4)	(9.4)	(8.3)	(6.6)	(7.0)	(6.6)	(8.7)	(9.4)
K	15	15	10	14	34	68	102	108	69	75	72	51	41	85
	(17.9)	**(20.3)**	(7.9)	(6.7)	(6.9)	(7.7)	(6.4)	(6.1)	(5.3)	(6.2)	(8.1)	(7.0)	(8.5)	(7.2)
M	0	1	4	5	20	22	26	47	35	37	32	17	17	23
	(0.0)	(1.4)	(3.2)	(2.4)	(4.1)	(2.5)	(1.6)	(2.7)	(2.7)	(3.1)	(3.6)	(2.3)	(3.5)	(2.0)
F	10	4	8	9	24	28	48	43	38	65	40	39	16	48
	(11.9)	(5.4)	(6.3)	(4.3)	(4.9)	(3.2)	(3.0)	(2.4)	(2.9)	(5.4)	(4.5)	(5.3)	(3.3)	(4.1)
P	2	1	1	5	11	22	57	62	47	30	24	17	10	18
	(2.4)	(1.4)	(0.8)	(2.4)	(2.2)	(2.5)	(3.6)	(3.5)	(3.6)	(2.5)	(2.7)	(2.3)	(2.1)	(1.5)
S	1	3	11	9	27	38	72	78	56	57	40	48	24	65
	(1.2)	(4.1)	(8.7)	(4.3)	(5.5)	(4.3)	(4.5)	(4.4)	(4.3)	(4.7)	(4.5)	(6.6)	(5.0)	(5.5)
T	19	8	9	11	25	49	93	97	70	68	55	42	32	84
	(22.6)	**(10.8)**	(7.1)	(5.2)	(5.1)	(5.5)	(5.8)	(5.5)	(5.4)	(5.6)	(6.2)	(5.8)	(6.7)	(7.1)
W	0	0	0	1	4	15	30	37	41	27	19	12	12	11
	(0.0)	(0.0)	(0.0)	(0.5)	(0.8)	(1.7)	(1.9)	(2.1)	(3.1)	(2.2)	(2.1)	(1.6)	(2.5)	(0.9)
Y	2	2	4	4	18	37	29	55	56	44	49	32	18	44
	(2.4)	(2.7)	(3.2)	(1.9)	(3.7)	(4.2)	(1.8)	(3.1)	(4.3)	(3.7)	(5.5)	(4.4)	(3.7)	(3.7)
V	7	4	3	10	40	61	112	119	95	74	44	40	32	105
	(8.3)	(5.4)	(2.4)	(4.8)	(8.2)	(6.9)	(7.0)	(6.8)	(7.3)	(6.1)	(4.9)	(5.5)	(6.7)	(8.9)

Table 4 Frequency and percentage distribution of the value of stability change for 20 types of residues from window size 55

Residue	ΔΔG range (kcal mol^{-1})													
	>=3.0	2.5∼3.0	2.0∼2.5	1.5∼2.0	1.0∼1.5	0.5∼1.0	0.0∼0.5	(−0.5) ∼ 0.0	(−1.0) ∼ (−0.5)	(−1.5) ∼ (−1.0)	(−2.0) ∼ (−1.5)	(−2.5) ∼ (−2.0)	(−3.0) ∼ (−2.5)	<(−3.0)
A	35	34	74	154	378	675	1275	1360	900	786	561	434	270	823
	(6.1)	(6.4)	(7.7)	(8.9)	(9.5)	(9.4)	(9.7)	(9.4)	(8.7)	(8.4)	(8.1)	(7.4)	(7.3)	(8.7)
R	14	23	47	85	187	397	743	884	570	540	409	329	201	562
	(2.5)	(4.3)	(4.9)	(4.9)	(4.7)	(5.5)	(5.7)	(6.1)	(5.5)	(5.8)	(5.9)	(5.6)	(5.4)	(6.0)
N	30	32	46	77	207	396	762	881	610	496	369	328	217	508
	(5.3)	(6.0)	(4.8)	(4.5)	(5.2)	(5.5)	(5.8)	(6.1)	(5.9)	(5.3)	(5.3)	(5.6)	(5.8)	(5.4)
D	33	27	32	84	213	422	785	876	639	578	443	361	239	553
	(5.8)	(5.1)	(3.3)	(4.9)	(5.4)	(5.8)	(6.0)	(6.0)	(6.1)	(6.2)	(6.4)	(6.1)	(6.4)	(5.9)
C	5	2	9	21	55	133	272	380	205	129	111	122	64	167
	(0.9)	(0.4)	(0.9)	(1.2)	(1.4)	(1.8)	(2.1)	(2.6)	(2.0)	(1.4)	(1.6)	(2.1)	(1.7)	(1.8)
Q	11	17	41	90	182	310	553	581	439	333	253	209	129	361
	(1.9)	(3.2)	(4.3)	(5.2)	(4.6)	(4.3)	(4.2)	(4.0)	(4.2)	(3.6)	(3.6)	(3.6)	(3.5)	(3.8)
E	52	33	78	138	268	437	715	764	562	599	423	350	218	461
	(9.1)	(6.2)	(8.2)	(8.0)	(6.8)	(6.1)	(5.5)	(5.3)	(5.4)	(6.4)	(6.1)	(5.9)	(5.9)	(4.9)
G	51	49	82	130	330	565	1029	1116	807	705	572	467	293	769
	(8.9)	(9.2)	(8.6)	(7.5)	(8.3)	(7.8)	(7.8)	(7.7)	(7.8)	(7.5)	(8.2)	(7.9)	(7.9)	(8.1)
H	14	11	20	42	68	132	211	241	151	157	113	93	63	110
	(2.5)	(2.1)	(2.1)	(2.4)	(1.7)	(1.8)	(1.6)	(1.7)	(1.5)	(1.7)	(1.6)	(1.6)	(1.7)	(1.2)
I	16	28	43	88	195	367	714	730	531	494	378	325	213	450
	(2.8)	(5.3)	(4.5)	(5.1)	(4.9)	(5.1)	(5.4)	(5.0)	(5.1)	(5.3)	(5.4)	(5.5)	(5.7)	(4.8)
L	23	41	80	144	335	638	1054	1163	840	765	552	418	312	751
	(4.0)	(7.7)	(8.4)	(8.3)	(8.5)	(8.8)	(8.0)	(8.0)	(8.1)	(8.2)	(7.9)	(7.1)	(8.4)	(8.0)
K	68	61	83	131	248	538	925	954	692	736	503	467	306	691
	(11.9)	**(11.5)**	(8.7)	(7.6)	(6.3)	(7.5)	(7.1)	(6.6)	(6.7)	(7.9)	(7.2)	(7.9)	(8.2)	(7.3)
F	30	16	31	41	112	187	415	420	307	339	230	190	108	281
	(5.3)	(3.0)	(3.2)	(2.4)	(2.8)	(2.6)	(3.2)	(2.9)	(3.0)	(3.6)	(3.3)	(3.2)	(2.9)	(3.0)
P	26	24	33	40	118	216	406	420	396	351	248	224	138	386
	(4.6)	(4.5)	(3.5)	(2.3)	(3.0)	(3.0)	(3.1)	(2.9)	(3.8)	(3.7)	(3.6)	(3.8)	(3.7)	(4.1)
S	18	24	57	84	219	347	629	801	595	435	369	333	206	527
	(3.2)	(4.5)	(6.0)	(4.9)	(5.5)	(4.8)	(4.8)	(5.5)	(5.7)	(4.6)	(5.3)	(5.7)	(5.5)	(5.6)
T	48	35	66	141	261	402	776	829	604	580	430	379	237	622
	(8.4)	(6.6)	(6.9)	(8.2)	(6.6)	(5.6)	(5.9)	(5.7)	(5.8)	(6.2)	(6.2)	(6.4)	(6.4)	(6.6)
Y	38	27	39	56	153	272	451	552	461	491	351	278	176	446
	(6.7)	(5.1)	(4.1)	(3.2)	(3.9)	(3.8)	(3.4)	(3.8)	(4.4)	(4.5)	(5.0)	(4.7)	(4.7)	(4.7)
V	40	27	57	108	258	455	887	943	748	612	411	375	216	674
	(7.0)	(5.1)	(6.0)	(6.3)	(6.5)	(6.3)	(6.8)	(6.5)	(7.2)	(6.5)	(5.9)	(6.4)	(5.8)	(7.1)

Table 5 Sequence triplets and corresponding frequency distribution of the value of stability change

ΔΔG range (kcal mol^{-1})	Sequence triplet									
	A*I	L*L	N*F	G*W	K*E	A*N	F*V	L*A	C*E	P*V
>=3.0	4	0	0	0	0	0	7	0	0	0
1.5∼3.0	10	4	0	0	1	0	2	2	6	1
0.0∼1.5	10	23	10	1	20	16	16	14	17	11
(−1.5) ∼ 0.0	19	20	16	23	12	15	3	9	2	12
(−3.0) ∼ (−1.5)	10	3	24	12	1	0	0	1	0	1
<(−3.0)	12	9	4	0	1	0	0	0	0	0
Total	65	59	54	36	35	31	28	26	25	25
Occurrence (%)	3.5	3.2	2.9	1.9	1.9	1.7	1.5	1.4	1.3	1.3
Stabilizing (%)	36.9	45.8	18.5	2.8	60.0	51.6	**89.3**	61.5	**92.0**	48.0
Destabilizing (%)	63.1	54.2	**81.5**	**97.2**	40.0	48.4	10.7	38.5	8.0	52.0
Log-odds ratio	−0.23	−0.07	−0.64	−1.54	0.18	0.03	0.92	0.20	1.06	−0.03

The mutants with positive ΔΔG are considered as stabilizing and that with negative ΔΔG are considered as destabilizing as per the conventions used in ProTherm database (27,28).

In Table 3, interestingly, it seems that the positive stability change (stabilizing) gives a greater specificity. For instance, the residue Leu has a high percentage of occurrences than the others when the range of stability change is between 0.5 and 2.5 kcal mol^{-1}; the residue Lys and Thr also have a high percentage when the value of stability change is larger than 2.5 kcal mol^{-1}. Focusing on the combination of residues, the residues Lys, Phe and Thr are significantly represented when the value of stability change is larger than 3 kcal mol^{-1}. Unusually, the residue Ala appears frequently when stabilizing as well as destabilizing. Comparing to Table 4, the residue Lys still gives the same specificity. However, the other ones become insignificant in occurrence percentage. The reason could be that the compositional specificity of residues is lost when larger window size introduces a lot of irrelevant residues. From the results described above, it explores the possibility of the compositional specificity of residues in different ranges of stability change.

Sequence analysis of pattern finding

For discovering the order specificity, we collected a group of sequence fragments comprised of triplets and quintets, which appear most frequently on the dataset. Table 5 lists the topmost 10 frequent triplets regardless of the central mutated residue (represented by *) out of a total of 293 triplets in the dataset. By dividing those triplets into different ranges of stability change, it is interesting that some triplets exist mostly in particular ranges. For instance, one of the most frequently occurring triplets G*W destabilizes the protein in 97% of the mutants and the log-odd ratio is −1.54. Besides, triplet N*F appears mostly

Table 6 Sequence quintets and corresponding frequency distribution of the value of stability change

ΔΔG range (kcal mol^{-1})	Sequence quintet									
	MN*FE	AL*LG	TG*WD	TF*VT	CP*VY	RC*EL	AQ*AG	AK*EL	DA*IK	AQ*LG
>=3.0	0	0	0	7	0	0	0	0	4	0
1.5∼3.0	0	0	0	2	1	6	0	0	7	0
0.0∼1.5	9	21	1	15	11	17	3	16	5	0
(−1.5) ∼ 0.0	12	15	23	2	12	1	20	5	5	18
(−3.0) ∼ (−1.5)	20	0	12	0	0	0	0	0	0	0
<(−3.0)	1	0	0	0	0	0	0	1	0	1
Total	42	36	36	26	24	24	23	22	21	19
Occurrence (%)	2.3	1.9	1.9	1.4	1.3	1.3	1.2	1.2	1.1	1.0
Stabilizing (%)	21.4	58.3	2.8	**92.3**	50.0	**95.8**	13.0	72.7	76.2	0.0
Destabilizing (%)	78.6	41.7	**97.2**	7.7	50.0	4.2	**87.0**	27.3	23.8	**100.0**
Log-odds ratio	−0.56	0.15	−1.54	1.08	0.00	1.36	−0.82	0.43	0.51	-

The mutants with positive ΔΔG are considered as stabilizing and that with negative ΔΔG are considered as destabilizing as per the conventions used in ProTherm database (27,28).

in a ΔΔG range of −3 to 0 kcal mol^{-1}. By contrast, triplets F*V and C*E stabilize the protein in 89.3% and 92.0% of the mutants, respectively. The log-odd ratio for the mutants with the pattern of F*V and C*E are, respectively, 0.92 and 1.06.

Extending the observation to 4 neighbors, the topmost 10 frequent quintets are listed in Table 6, in which the pattern TF*VT and RC*EL stabilize the protein. The log-odd ratios for these patterns of mutants are 1.08 and 1.36, respectively. Quintets TG*WD and AQ*AG destabilize the protein in 97.2% and 87.0% of the mutants, respectively, and the ΔΔG values are in range of −3 to 0 kcal mol^{-1}. The respective log-odd ratios are −1.54 and −0.82 for the patterns TG*WD and AQ*AG. Interestingly, quintet AQ*LG only destabilize the protein. The results imply the existence of certain patterns which can help both understanding and predicting the stability change of protein mutants.

Exploring important factors and decision rules

Besides applying the sequence analysis, iPTREE-2 can also play a role in mining knowledge of predicting protein

Table 7 Decision rules built from S1859 and corresponding details of prediction performance

No.	Antecedent	Predicted value (kcal mol^{-1})	Rule size	STD (kcal mol^{-1})	MAE (kcal mol^{-1})	Number of data	Data percentage
1	Md [Y W V R P M L I G F C] Mi [T S P K H G A]	−2.049	2	2.17	1.57	469	25.2
2	Md [T S Q N K H E D A] Mi [W T S R Q P N K H G E D C A] W_1 [T S R Q P M K I G F D A -] W_3 [W T S P M L K H F E D C] W_4 [V T S Q N L I H G E C A -] W_6 [W V S R Q P N L K I H G F E D A -] N_Q<1.5	−0.110	7	0.85	0.61	177	9.5
3	Md [W R L I F C] Mi [Y W V R Q N M L I F E D C] W_4 [Y V R Q P N M I F E C A]	−0.906	3	1.30	0.93	148	8.0
4	Md [Y V P M G] Mi [Y W V R Q N M L I F E D C] W_4 [Y V R Q P N M I F E C A] W_6 [W T R Q P M I H F D C]	−0.400	4	1.06	0.77	89	4.8
5	Md [T S Q N K H E D A] Mi [W T S R Q P N K H G E D C A] W_1 [Y V N L H E C] W_3 [W T S P M L K H F E D C] W_4 [V T S Q N L I H G E C A -] W_6 [W V S R Q P N L K I H G F E D A -] N_Q<1.5	−0.771	7	1.07	0.81	71	3.8
6	Md [T S Q N K H E D A] Mi [S R Q N M H E D C A] W_2 [Y W V T S R N G F E D A] W_4 [Y W R P M K F D] W_5 [Y W V T S R K I H G E D] W_6 [V T R P N I F E A]	−0.906	6	0.85	0.61	64	3.4
7	Md [T S Q N K H E D A] Mi [W T S R Q P N K H G E D C A] W_1 [T S R Q P M K I G F D A -] W_3 [W T S P M L K H F E D C] W_4 [V T S Q N L I H G E C A -] W_6 [W V S R Q P N L K I H G F E D A -] N_Q >=1.5	0.633	7	0.84	0.66	48	2.6
	Mean		5	1.16	0.85	152	8.2

The mutants with positive ΔΔG are considered as stabilizing and that with negative ΔΔG are considered as destabilizing as per the conventions used in ProTherm database (27,28).

STD: standard deviation

MAE: mean absolute error

stability change. To achieve a balance between lifting prediction performance and simplifying tree models, the variable set we used is based on CO-7. When building the decision tree model, the variable that is selected earlier implies it is impact to the built tree. As a result, we found that the mutation residue in wild type and the mutation residue in mutant protein were the two most important variables among those input variables. Interestingly, it agrees with our previous study [20], revealing the two factors and the temperature make a major contribution to the distinct ability of predicting protein stability change. The prediction and analysis upon amino acid substitutions with 48 amino acid properties have also been presented previously [26].

When completing the decision tree model, it can transform to a rule base that consists of decision rules. For the purposes of demonstration, the maximum depth of the decision tree model was limited to 7 to reduce the rule size that means the length of antecedents of rules. Nevertheless, the free data mining software [32] for the implement of regression tree is available in the Internet. Table 7 shows the rules with larger data number that refers to the number of samples to which a rule can be applied in the dataset. Whereas a rule with large data number may serve a general phenomenon, we focus on the first rule:

If the deleted-residue belongs to Y, W, V, R, P, M, L, I, G, F or C, and the introduced-residue belongs to T, S, P, K, H, G or A, then the predicted stability change value is −2.05 kcal mol^{-1}.

Interestingly, the previous study has revealed some properties about Pro and Ala in protein stability. The proline-free triple mutant P7A/P9A/P50A was investigated using Fourier-transform infrared (FTIR) spectroscopy [33]. The thermal stability of the proline-free mutant is reduced by 15 °C as compared to the wild type. Moreover, the rule is generally compatible with our previous study [20], describing

If temperature is between 4 °C and 40 °C, and the introduced residue is Alanine, then the predicted stability change will be negative.

This first rule with less rule size of 2 covers a large number of samples of 469 that accounts for 25.2% of the whole dataset. However, the high standard deviation (STD=2.17 kcal mol^{-1}) indicates the covered samples are a little spread so that the confidence in prediction may descend. By contrast, the sixth rule:

If the deleted-residue belongs to T, S, Q, N, K, H, E, D or A; the introduced-residue belongs to S, R, Q, N, M, H, E, D, C or A; the second neighbor toward N-terminus belongs to Y, W, V, T, S, R, N, G, F, E, D or A; the three neighbors toward C-terminus belongs to [Y W R P M K F D], [Y W V T S R K I H G E D] and [V T R P N I F E A], respectively; then the predicted stability change value is −0.91 kcal mol^{-1}.

performs a better standard deviation (STD=0.85 kcal mol^{-1}) and a lower mean absolute error (MAE=0.61 kcal mol^{-1}) with increasing rule size of 6 and decreasing data number of 64, which implies that the precise rules may improve the prediction performance but narrow the applied scope.

The seven listed rules can cover a total 57.3% of data, while the numbers of destabilizing and stabilizing samples are 778 and 288, respectively. The major reason is due to the percentage of two kinds of training data. The rules are deduced from the dataset to represent them as much as possible and the covered samples usually exhibit the similar characteristic of data distribution. Those interpretable rules may be consistent with previous reports or lead to a new discovery that still requires confirmation. However, those derived rules presenting an understandable concept of dataset can be validated further to be usable knowledge.

Conclusions

In this paper, the proposed iPTREE-2 has been effectively applied to a large database of the thermodynamic data of mutant proteins to predict the value of protein stability change. By tuning pruning level, the correlation coefficient between predicted and experimental values can successfully reach to 0.70 when based on sequence information. Through comparison of designed variables sets, selecting a small window size as input is acceptable and efficient when based on sequence information including both composition and order. In contrast to sequence information including composition without order, choosing a larger window size is helpful to lift up the prediction performance.

From observation of different scales of window size, the compositional specificity of residues really exists in different ranges of stability change. The found triplets and quintets can offer a possibility of both understanding and predicting the stability change of protein mutants. iPTREE-2 also provided a mechanism for mining knowledge of predicting protein stability change. The results reveal that some factors play an important role and the built interpretable rules can be validated by well-known researches, and all the knowledge can provide us with more understanding about the protein stability change.

Future, since a set of significant features is helpful for the establishment of more interpretable rules, as well as the

prediction performance. Thus, selecting appropriate variables will be a worthy issue.

References

1. Daggett V, Fersht AR (2003) Trends Biochem Sci 28:18–25
2. Saven JG (2002) Curr Opin Struct Biol 12:453–458
3. Mendes J, Guerois R, Serrano L (2002) Curr Opin Struct Biol 12:441–446
4. Bolon DN, Marcus JS, Ross SA, Mayo SL (2003) J Mol Biol 329:611–622
5. Looger LL, Dwyer MA, Smith JJ, Hellinga HW (2003) Nature 423:185–190
6. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A (1999) Protein Eng 12:549–555
7. Guerois R, Nielsen JE, Serrano L (2002) J Mol Biol 320:369–387
8. Prevost M, Wodak SJ, Tidor B, Karplus M (1991) Proc Natl Acad Sci USA 88:10880–10884
9. Gilis D, Rooman M (1997) J Mol Biol 272:276–290
10. Parthiban V, Gromiha MM, Schomburg D (2006) Nucleic Acids Res 34:W239–W242
11. Funahashi J, Takano K, Yutani K (2001) Protein Eng 14:127–134
12. Capriotti E, Fariselli P, Casadio R (2004) Bioinformatics 20 Suppl 1:I63–I68
13. Capriotti E, Fariselli P, Casadio R (2005) Nucleic Acids Res 33: W306–W310
14. Cheng J, Randall A, Baldi P (2006) Proteins 62:1125–1132
15. Xiong W, Wang JTL, Shasha D, Shapiro BA, Rigoutsos I, Kaizhong Z (2002) Knowledge and Data Engineering, IEEE Transactions on 14:731–749
16. Creighton C, Hanash S (2003) Bioinformatics 19:79–86
17. Oyama T, Kitano K, Satou K, Ito T (2002) Bioinformatics 18:705–714
18. Baldi P, Brunak S (2001) Bioinformatics: the machine learning approach. MIT Press, Cambridge, Mass
19. Larose DT (2005) Discovering knowledge in data: An introduction to data mining. Wiley-Interscience, Hoboken, New York
20. Huang LT, Gromiha MM, Hwang SF, Ho SY (2006) Computational Biology and Chemistry 30:408–415
21. Bordner AJ, Abagyan RA (2004) Proteins 57:400–413
22. Casadio R, Compiani M, Fariselli P, Vivarelli F (1995) Proc Int Conf Intell Syst Mol Biol 3:81–88
23. Frenz CM (2005) Proteins 59:147–151
24. Lacroix E, Viguera AR, Serrano L (1998) J Mol Biol 284: 173–191
25. Munoz V, Serrano L (1997) Biopolymers 41:495–509
26. Huang LT, Saraboji K, Ho SY, Hwang SF, Ponnuswamy MN, Gromiha MM (2007) Biophysical Chemistry 125:462–470
27. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A (2004) Nucleic Acids Res 32:D120–D121
28. Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Sarai A (1999) Nucleic Acids Res 27:286–288
29. Breiman L (1984) Classification and regression trees. Wadsworth International Group, Belmont, CA
30. Bai JP, Utis A, Crippen G, He HD, Fischer V, et al (2004) J Chem Inf Comput Sci 44:2061–2069
31. Deconinck E, Zhang MH, Coomans D, Vander Heyden Y (2006) J Chem Inf Model 46:1410–1419
32. Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco
33. Zscherp C, Aygun H, Engels JW, Mantele W (2003) Biochim Biophys Acta 1651:139–145